Journal of Applied Mathematics and Stochastic Analysis 5, Number 3, Fall 1992, 237-260

ON A MULTILEVEL CONTROLLED BULK QUEUEING SYSTEM $M^X/G^{r,R}/1$ [1]

LEV ABOLNIKOV
Department of Mathematics, Loyola Marymount University
Los Angeles, CA 90045, USA

JEWGENI H. DSHALALOW
Department of Applied Mathematics, Florida Institute of Technology
Melbourne, FL 32901, USA

ABSTRACT

The authors introduce and study a class of bulk queueing systems with a compound Poisson input modulated by a semi-Markov process, multilevel control service time and a queue length dependent service delay discipline. According to this discipline, the server immediately starts the next service act if the queue length is not less than r; in this case all available units, or R (capacity of the server) of them, whichever is less, are taken for service. Otherwise, the server delays the service act until the number of units in the queue reaches or exceeds level r.

The authors establish a necessary and sufficient criterion for the ergodicity of the embedded queueing process in terms of generating functions of the entries of the corresponding transition probability matrix and of the roots of a certain associated functions in the unit disc of the complex plane. The stationary distribution of this process is found by means of the results of a preliminary analysis of some auxiliary random processes which arise in the "first passage problem" of the queueing process over level r. The stationary distribution of the queueing process with continuous time parameter is obtained by using semi-regenerative techniques. The results enable the authors to introduce and analyze some functionals of the input and output processes via ergodic theorems. A number of different examples (including an optimization problem) illustrate the general methods developed in the article.

Key words: Queueing process, modulated random measure, semi-Markov process, semi-regenerative process, embedded Markov chain, semi-Markov modulated marked Poisson process, equilibrium, continuous time parameter process.

AMS Subject Classification: Primary 60K10, 60K15, 60K25, Secondary 90B22, 90B25.

[1]Received: December, 1991. Revised: July, 1992.

1. INTRODUCTION AND SUMMARY OF PREVIOUS WORK

A multilevel control strategy in a bulk queueing system is based on the utilization of certain feedback relationships between parameters of both bulk arrival and bulk service processes and a current number of units in the system (or in the queue). Using this control strategy it is possible, for example, to respond to an excessively long (or, conversely, too short) queue by changing the rates of the arrival and service processes, or by changing the sizes of arriving groups of units or groups taken for service.

Another possibility to control the queueing process in bulk queueing systems is an assumption that the server of capacity R can delay a new service act until the number of units in the queue reaches a certain control level r (where $r \leq R$). A "service delay discipline" of this type may be useful in reducing start-up costs and, in combination with a multilevel control strategy, offers a considerable scope for improvements and optimization of the queueing process.

Special cases of a queueing system with service delay discipline appeared in Chaudhry and Templeton [7] and Chaudhry, Madill and Briére [8] without bulk input. The authors called them queueing systems with a *quorum* and denoted $M/G^{a,b}/1$. A more general system was introduced and studied even earlier by Neuts [23], where, in the version of $M/G^{a,b}/1$ treated in [8], the author of [23] additionally assumed that service times may depend on groups sizes (between a and b) taken for service.

A queueing system $M/G^{a,b}/1$ becomes more attractive (both theoretically and practically) if, in addition to the assumptions about quorum systems accepted in [7,8,23], the input stream is allowed to be bulk. In contrast with models of type $M/G^{a,b}/1$, in which the queue length, being less than a, will always (with probability 1) reach a, in systems $M^X/G^{a,b}/1$ (with general bulk input) the probability of reaching exactly level a in similar situations is less than 1 (in most cases even very small). Or more precisely, the queue length can exceed this level with a positive probability by any conceivable integer value. Behaviors of such processes about some critical level generally differs from those of processes with continuous state space which is treated by various methods called "level crossing analysis" [see 5,10,24]. Consequently, this fact makes a preliminary analysis of a corresponding first passage problem (of the queueing process over some fixed level) a necessary and essential part of the treatment of the system. As a separate kind of the first passage problem, this was recently studied in Abolnikov and Dshalalow [3].

Another considerable generalization of the $M/G^{a,b}/1$ system, employed in this article, is the assumption that interarrival times of the input stream, the sizes of arriving batches of customers, and service time distributions of groups of customers taken for service, all depend upon

the queue length (multilevel control policy). This essentially enlarges a class of real-world systems to which the results obtained are applicable.

It is necessary to mention that unlike a model studied in [8] the authors of this paper do not assume the dependence of service on group sizes. [The latter option, however, was included and extensively studied in Dshalalow [14] and [15], and Dshalalow and Tadj [17-19], and it may be combined with all options set in the present paper.] Other special cases of this model were considered in the literature on queueing theory. An idea to employ a multilevel control policy in bulk queueing systems, presumably, belongs to Bachary and Kolesar [6]. However, methods used there suffer from insufficient analytical justification. Some special cases of a multilevel control strategy with respect to more general queueing systems and inventory and dam models were considered in [1]. An example of an application of a multilevel control strategy to a general bulk queueing system (but with no service delay discipline) is contained in [12].

The main purpose of this article is to develop a general mathematical model which would take into account mentioned above features of a queueing system with a bulk input, batch service, multilevel control strategy and a queue length dependent service delay discipline.

In the present article a general bulk queueing system with a multilevel control and a queue length dependent service delay discipline is studied. The results obtained in the paper generalize, compliment or refine similar results existing in the literature [2,6-8,11,12,21,23].

In the first section the authors give a formal definition of a multilevel control bulk queueing system $M^X/G^{r,R}/1$ (with a queue length dependent service delay discipline). In order to describe rigorously the input to the system, the authors use the general notion of a modulated random measure recently introduced and studied in Dshalalow [13]. The authors establish necessary and sufficient conditions for the ergodicity of the queueing process with discrete and continuous time parameters and study its steady state distributions in both cases. A recently developed new analysis of a class of general Markov chains in Abolnikov and Dukhovny [4] is used. Due to the queue length dependent service delay discipline assumed in the article, an auxiliary random process describing the value of the first excess of the queue length above a certain control level appears to be one of the kernel components in the analysis of the queueing process (section 3). Using this random process, the authors derive the invariant probability measure of an embedded process in terms of generating functions and the roots of a certain associated function in the unit disc of the complex plane (section 4). The stationary distribution of the queueing process with continuous time parameter is obtained by using semi-regenerative techniques (section 6). The results of this section together with [13] enable the authors to intro-

duce and study some functionals of the input and output processes via ergodic theorems. A number of different examples (including an optimization problem) illustrate the general methods developed in the article.

2. FORMAL DESCRIPTION OF THE SYSTEM

We begin this section with the definition of the modulated random measure introduced and studied in Dshalalow [11] (formulated there for a more general case). All stochastic processes below will be considered on a common probability space $\{\Omega, \mathcal{F}, (P^x)_{x\in\Psi}\}$, with $\Psi = \{0,1,...\}$.

2.1 Definition. Let $E = \mathbb{R}_+$ with its natural topology and let $\mathcal{M}$ be the space of all Radon measures on the Borel σ-algebra $\mathcal{B}(E)$. Denote $\mathcal{C}_K$ the space of all continuous functions on E with compact support and denote $\mathfrak{M}$ the Baire σ-algebra in $\mathcal{M}$ generated by all maps $\mu \mapsto \int f d\mu$, $f \in \mathcal{C}_K$.

(*i*) Let $\{\Omega, \mathcal{F}, P, \xi(t),\ t \in E\} \to \Psi$ be a stochastic process on E and let ξ_ω denote the ω-section of ξ. Then for $\Gamma \subseteq \Psi$ and $B \in \mathcal{B}(E)$ we define $Y_\Gamma = B \cap \xi_\omega^{-1}(\Gamma)$ and call it the *holding time of ξ in Γ on set B*. For each fixed ω, Y_Γ is a measurable subset of E which can be measured by any Radon measure on $\mathcal{B}(E)$. In general, Y_Γ is a mapping from Ω into $\mathcal{B}(E)$ which can be made a *random set* after we define the σ-algebra $\{\Lambda \subseteq \mathcal{B}(E): Y_\Gamma^{-1}(\Lambda) \in \mathcal{F}\}$.

(*ii*) Consider for each j a marked random measure $Z_j = \sum_i S_{ij}\varepsilon_{\tau_{ij}}$ (where ε_x denotes the Dirac point mass) with mark space $\{0,1,...\}$ and introduce

$$Z^\xi = Z^\xi(\omega, B) = \sum_{j=0}^{\infty} Z_j(Y_{\{j\}}).$$

The random measure $Z^\xi: (\Omega, \mathcal{F}) \to (\mathcal{M}, \mathfrak{M})$ is called a *marked random measure modulated by the process ξ*. The marked random measure Z^ξ can be more vividly represented in the form

$$Z^\xi = \sum_{n=1}^{\infty} S_{n\xi}\varepsilon_{\tau_{n\xi}}. \qquad \square$$

Let $\{Q(t);\ t \geq 0\} \to \Psi = \{0,1,...\}$ be a stochastic process describing the number of units at time t in a single-server queueing system with an infinite waiting room. Following the introduction, $\{T_n; n \in \mathbb{N}_0, T_0 = 0\}$ is the sequence of successive completions of service and $Q_n = Q(T_n + 0)$.

Input Process. Let $C = \sum_{n=0}^{\infty} \varepsilon_{T_n}$. Define $\xi(t) = Q(T_{C([0,t])} + 0)$, $t \geq 0$. Then the input is a compound Poisson process *modulated by ξ* according to definition 2.1, from which it follows that customers arrive at random instants of time $\tau_{n\xi}$, $n = 1,2,...$, that form a point process modulated by ξ with $S_{i\xi(t)} - S_{i-1,\xi(t)} = X_{i\xi(t)}$ as the ith batch size of the input flow. Thus, in our

case $\{X\} = \{X_{i\xi(t)}\}$ is an integer-valued doubly stochastic sequence describing the sizes of groups of entering units. We assume that given $\xi(t)$ all terms of $\{X\}$ are independent and identically distributed. Denote $a_{\xi(t)}(z) = E[z^{X_{i\xi(t)}}]$, the generating function of ith component of the process $\{X\}$, with $\alpha_{\xi(t)} = E[X_{i\xi(t)}] < \infty$, $i = 1,2,\ldots$.

Service Time and Service Discipline. If at time $T_n + 0$ the queue length, Q_n, is at least r (a positive integer number less than or equal to R), the server takes a batch of units of size R (a positive integer denoting the capacity of the server) from the queue and then serves it during a random length of time σ_{n+1}. [We assume that σ_{n+1} has a probability distribution function $B_{Q_n} \in \{B_0, B_1, \ldots\}$, where the latter is a given sequence of arbitrary distribution functions with finite means $\{b_0, b_1, \ldots\}$.] Otherwise, the server idles until the queue length for the first time reaches or exceeds the level r. Let $\gamma_n = inf\{k \in \mathbb{N}: \tau_{k\xi} \geq T_n\}$, $n \in \mathbb{N}_0$. Then the size of the first group after T_n (which arrives at instant of time $\tau_{\gamma_n \xi}$) is X_{γ_n, Q_n}.

For more convenience in notation, we reset the first index-counter of the process $\{X\}$ on 1 after time t hits T_n. Therefore, in the light of the new notation, $X_{1Q_n}, X_{2Q_n}, \ldots$ are the sizes of successive groups of units arriving at the system after T_n. Let $S_{kn} = X_{0Q_n} + X_{1Q_n} + \ldots + X_{kQ_n}$, where $X_{0Q_n} = Q_n$. Then, given Q_n, $\{S_{kn};\ k \in \mathbb{N}_0\}$ is an integer-valued delayed renewal process. Denote $\nu_n = inf\{k \geq 0:\ S_{kn} \geq r\}$ the random index when the process $\{S_{kn}\}$ first reaches or exceeds level r given that the queue length is Q_n. Thus, $\tau_{\nu_n \xi}$ is the *instant of the first excess of level* r by the queueing process after time T_n (more accurate but less friendly notation for this instant would be $\tau_{\gamma_n + \nu_n - 1, \xi}$).

For the next constructions we need a more universal notation for the instant of the first excess of level r, appropriate for the situations when this instant occurs after T_n or at T_n. In other words we define

$$\theta_n(Q_n) = \theta_n = \begin{cases} \tau_{\nu_n \xi}, & X_{0Q_n} = Q_n < r \\ T_n, & X_{0Q_n} = Q_n = S_{kn} \geq r. \end{cases} \tag{2.2}$$

At that instant of time the server is supposed to take a batch of the size $min\{Q(\theta_n), R\}$ for service. In other words, if $Q_n \geq r$, $T_{n+1} - T_n$ coincides with length of service time σ_{n+1} of the $n+1$st batch. If $Q_n < r$ the interval $(T_n, T_{n+1}]$ contains the waiting time for $X_{1Q_n} + \ldots + X_{\nu_n Q_n}$ units to arrive and the actual service time σ_{n+1}.

Finally, denoting $V_n = Z^{\xi}(\sigma_n)$ we can abbreviate the definition of the servicing process through the following relation for $\{Q_n\}$:

$$Q_{n+1} = \begin{cases} (S_{\nu_n} - R)^+ + V_{n+1}, & Q_n < r \\ (Q_n - R)^+ + V_{n+1}, & Q_n \geq r \,. \end{cases} \tag{2.3}$$

3. FIRST PASSAGE PROBLEM

In the following sections we will be using some basic results on the first passage problem stated and developed in Abolnikov and Dshalalow [3]. Some of the results of [3] (which the authors will highlight in this section) were obtained by Dynkin [20] and Takács [25] more general processes. However, for convenience in notation and with the purpose of a more specific terminology only results of [3] will be mentioned.

First we treat the process $\{S_{\nu_n}\}$ without any connection to the queueing system. For this reason we will temporarily simplify the notation introduced in section 2 by suspending the second subscript in the sequence $\{X\}$ and the corresponding index in the probability measure P^i and expectation E^i.

Therefore, in this section we will discuss the "critical behavior" of a compound Poisson process Z determined by a Poisson process $\tau = \{\tau_n = t_0 + t_1 + ... + t_n; n \geq 0,\ t_0 = 0\}$ on $\mathbb{R}_+$ marked by a discrete-valued delayed renewal process $S = \{S_n = X_0 + X_1 + ... + X_n; n \geq 0\}$ on Ψ. As mentioned above, we assume that the processes τ and S are independent. We also assume that inter-renewal times $t_n = \tau_n - \tau_{n-1}$, are described in terms of its common Laplace-Stieltjes transform $e(\theta) = E[e^{-\theta t_n}] = \frac{\lambda}{\lambda + \theta}$, $n = 1,2,...$.

For a fixed integer $r \geq 1$ we will be interested in the behavior of the process S and some related processes about level r.

The following terminology is introduced and will be used throughout the paper.

3.1 Definitions.

(*i*) For each n the random variable $\nu_n = inf\{k \geq 0: S_k \geq r\}$ (defined in the previous section) is called the *index of the the first excess* (above level $r-1$).

(*ii*) The random variable S_{ν_n} is called the *level of the first excess* (above $r-1$).

(*iii*) The random variable τ_{ν_n} is known as the *first passage time* of S of level r.

(*iv*) The random variable $\mathfrak{J}_n = S_{\nu_n} - S_0$ is called the *increment of the input process over the time interval* $[T_n, \theta_n]$, or shortly, the *total increment.*

Let

(3.1a) $$\gamma^{(i)}(\theta,z) = E^i[e^{-\theta\tau_{\nu_0}} z^{\nu_0}],\ \mathfrak{g}^{(i)}(\theta,z) = E^i[e^{-\theta\tau_{\nu_0}} z^{S_{\nu_0}}],$$

$$G_i(\theta,z) = \sum_{j\geq 0} E^i[e^{-\theta\tau_j} z^{S_j} I_{U_{r-1}}(S_j)],$$

where $U_p = \{0,1,\ldots,p\}$ and I_A is the indicator function of a set A. We call $G_i(\theta,z)$ the *generator of the first excess level.* We will also use the following functionals of marginal processes:

(3.1b) $$\gamma^{(i)}(z) = \gamma^{(i)}(0,z),$$

(3.1c) $$\mathfrak{g}^{(i)}(z) = \mathfrak{g}^{(i)}(0,z),$$

(3.1d) $$G_i(z) = G_i(0,z).$$

It is readily seen that $G_i(z)$ is a polynomial of $(r-1)$th degree.

We formulate the main theorems from Abolnikov and Dshalalow [3] and give formulas for the joint distributions of the first passage time and the random variables listed in 3.1 (*i-iii*).

3.2 Theorem. *The functional* $\gamma^{(i)}(\theta,z)$ (*of the first passage time and of the index of the first excess level*) *satisfies the following formula:*

(3.2a) $$\gamma^{(i)}(\theta,z) = \begin{cases} e(\theta) z \mathfrak{D}_x^{r-i-1}\left\{\dfrac{1-a(x)}{(1-x)(1-ze(\theta)a(x))}\right\}, & i<r \\ 1, & i\geq r, \end{cases}$$

where

(3.2b) $$\mathfrak{D}_x^k = \lim_{x\to 0} \frac{1}{k!}\frac{\partial^k}{\partial x^k},\ k\geq 0.$$

Specifically, the Laplace-Stieltjes transform of the first passage time, $\gamma^{(i)}(\theta,1)$, is as follows:

(3.2c) $$\gamma^{(i)}(\theta,1) = \begin{cases} e(\theta) \mathfrak{D}_x^{r-i-1}\left\{\dfrac{1-a(x)}{(1-x)(1-e(\theta)a(x))}\right\}, & i<r \\ 1, & i\geq r. \end{cases}$$

From formula (3.2a) we immediately obtain that the mean value of the index of the first excess equals

(3.3) $$\bar{\gamma}^{(i)} = \begin{cases} \mathfrak{D}_x^{r-i-1}\left\{\dfrac{1}{(1-x)[1-a(x)]}\right\}, & i<r \\ 0, & i\geq r. \end{cases}$$

From (3.2a) we also obtain the mean value of the first passage time:

(3.4) $$E^i[\tau_{\nu_0}] = \frac{1}{\lambda}\bar{\gamma}^{(i)}.$$

3.5 Theorem. *The generator* $G_i(\theta,z)$ *of the first excess level can be determined from the following formula:*

(3.5a)

$$G_i(\theta,z)=\begin{cases} z^i\mathfrak{D}_x^{r-i-1}\left\{\dfrac{1}{(1-x)[1-e(\theta)a(xz)]}\right\}, & i<r \\ 0, & i\geq r. \end{cases}$$

The rationale behind the use of the term "generator of the first excess level" comes from the following main result.

3.6 Theorem. *The functional* $\mathfrak{g}^{(i)}(\theta,z)$ *(of the first passage time and of the first excess level) can be determined from the formula*

(3.6a) $$\mathfrak{g}^{(i)}(\theta,z)=z^i-[1-e(\theta)a(z)]G_i(\theta,z).$$

3.7 Remark. To obtain the functionals of the marginal processes defined in (3.1b-3.1d) we set $e(\theta)=1$ in formulas (3.2a), (3.5a) and (3.6a).

3.8 Corollary. *The generating function* $\mathfrak{g}^{(i)}(z)$ *of the first excess level is determined by the following formula:*

(3.8a) $$\mathfrak{g}^{(i)}(z)=\begin{cases} z^i\mathfrak{D}_x^{r-i-1}\left\{\dfrac{a(z)-a(xz)}{(1-x)(1-a(xz))}\right\}, & i<r \\ z^i, & i\geq r. \end{cases}$$

By using change of variables in (3.8a) we can transform it into an equivalent expression

(3.8b) $$\mathfrak{g}^{(i)}(z)=\begin{cases} z^r\mathfrak{D}_x^{r-i-1}\left\{\dfrac{a(z)-a(x)}{(z-x)[1-a(x)]}\right\}, & i<r \\ z^i, & i\geq r. \end{cases}$$

3.9 Corollary.

(3.9a) $$\bar{\mathfrak{g}}^{(i)}=E^i[S_{\nu_0}]=i+\alpha\bar{\gamma}^{(i)},$$

where $\alpha=E[X_1]$.

Specifically, the mean value $\bar{\mathfrak{z}}^{(i)}=E^i[\mathfrak{z}_0]$ of the total increment is then

(3.9b) $$\bar{\mathfrak{z}}^{(i)}=\alpha\bar{\gamma}^{(i)}.$$

3.10 Remark. Now we notice that the above results can be applied to our queueing system, where in formulas (3.2a)-(3.9a) we supply $a(z)$ with subscript i.

4. EMBEDDED PROCESS $\{Q_n\}$

4.1 Definition. Let T be a stopping time for a stochastic process $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi},\ Q(t);\ t\geq 0\}\rightarrow(\Psi,\ \mathfrak{B}(\Psi))$. $\{Q(t)\}$ is said to have the *locally strong Markov property at* T if for each

bounded random variable $\zeta\colon \Omega \to \Psi^r$ and for each Baire function $f\colon \Psi^r \to \mathbb{R}$, $r = 1,2,...$, it holds true that

$$E^x[f \circ \zeta \circ \Theta_T \,|\mathfrak{F}_T] = E^{Z_T}[f \circ \zeta] \quad P^x\text{-a.s. on } \{T < \infty\},$$

where Θ_y is the shift operator.

From relation (2.3) and the nature of the input it follows that the process $\{\Omega, \mathfrak{F}, (P^x)_{x\in\Psi}, Q(t);\ t \geq 0\} \to \Psi = \{0,1,...\}$ possesses a *locally strong Markov property* at T_n, where T_n is a stopping time relative to the canonic filtering $\sigma(Q(y); y \leq t)$, $n = 1,2,...$. Thus the embedded process $\{\Omega, \mathfrak{F}, (P^x)_{x\in\Psi}, Q_n; n \in \mathbb{N}_0\} \to \Psi$ is a homogeneous Markov chain with transition probability matrix denoted by $A = (a_{ij})$.

4.2 Lemma. *The generating function $A_i(z)$ of ith row of matrix A can be determined from the following formulas:*

$$A_i(z) = K_i(z) H_i^{(r,R)}(z) \qquad \textit{where} \tag{4.2a}$$

$$H_i^{(r,R)}(z) = z^{-R}\mathfrak{g}^{(i)}(z) + \mathfrak{D}_y^{R-1}\left\{\frac{\mathfrak{g}^{(i)}(y) - z^{-R}\mathfrak{g}^{(i)}(yz)}{1-y}\right\},\ i \in \Psi, \tag{4.2b}$$

$$K_i(z) = \beta_i(\lambda_i - \lambda_i a_i(z)), \tag{4.2c}$$

$\beta_i(\theta)$, $Re(\theta) \geq 0$, is the Laplace-Stieltjes transform of the probability distribution function B_i, and $\mathfrak{g}^{(i)}$ satisfies one of the formulas (3.8a) *or* (3.8b), *taking into account remark* 3.10.

Proof. Since $A_i(z) = E^i[z^{Q_1}]$, formulas (4.2a) and (4.2b) follow from (2.3) and probability arguments similar to those in the proofs of section 3. Observe that since $\mathfrak{g}^{(i)}(z) = z^i$, $i \geq r$, in (3.8a), formula (4.2b) reduces to

$$H_i^{(r,R)} = z^{(i-R)^+},\ i \geq r, \tag{4.2d}$$

which also agrees with the result that could have been obtain directly from (2.3) for $i \geq r$. □

For analytical convenience and without considerable loss of generality we can drop the modulation of the input process and service control when the queue length exceeds a fixed (perhaps very large) level N. In other words, we assume that

(AS) $B_i(x) = B(x), \beta_i(\theta) = \beta(\theta), K_i(z) = K(z), b_i = b, \lambda(i) = \lambda_i = \lambda, a_i(z) = a(z), \alpha_i = \alpha,\ i > N,\ N \geq r-1$, where $\alpha_i = a_i'(1)$, $i \in \Psi$.

Given assumption (AS), it can be shown that the transition probability matrix A is reduced to a form of the $\Delta_{R,N}$-matrix introduced and studied in [4]. There the stochastic matrix $A = (a_{ij};\ i,j \in \Psi = \{0,1,...\})$ is called a $\Delta_{R,N}$-*matrix* if it is of the form

$$A = (a_{ij}\colon i,j \in \Psi\colon\ a_{ij} = k_{j-i+r},\ i > N,\ j \geq i-R;\ a_{ij} = 0,\ i > N,\ j < i-R),$$

where $\sum_{j=0}^{\infty} k_j \varepsilon_i$ is an atomic probability measure. The following two theorems are necessary to obtain all main results in this section.

4.3 Theorem (Abolnikov and Dukhovny [4]). *Let $\{Q_n\}$ be an irreducible aperiodic Markov chain with the transition probability matrix A in the form of a $\Delta_{r,N}$-matrix. Q_n is recurrent-positive if and only if*

$$\frac{d}{dz} A_i(z)\Big|_{z=1} < \infty,\ i = 0,1,...,N, \tag{4.3a}$$

and

$$\frac{d}{dz} K(z)\Big|_{z=1} < R. \tag{4.3b}$$

4.4 Theorem (Abolnikov and Dukhovny [4]). *Under the condition of* (4.3b) *the function $z^r - K(z)$ has exactly r roots that belong to the closed unit ball $\bar{B}(0,1)$. Those of the roots lying on the boundary $\partial B(0,1)$ are simple.*

Condition (4.3a) is obviously met and condition (4.3b) is equivalent to

$$\rho = \lambda\alpha b < R. \tag{4.5}$$

Therefore, given that $\rho < R$, the Markov chain $\{Q_n\}$ is ergodic. Let $P = (p_x; x \in \Psi)$ be the invariant probability measure of operator A and let $P(z)$ be the generating function of the components of vector $\boldsymbol{P}$. Now we formulate the main result of this section.

4.6 Theorem. *The embedded queueing process Q_n is ergodic if and only if $\rho < R$. Under this condition, $P(z)$ is determined by the following formula:*

$$P(z) = \frac{\sum_{i=0}^{N} p_i \left\{ z^R K_i(z) H_i^{(r,R)}(z) - z^i K(z) \right\}}{z^R - K(z)}, \tag{4.6a}$$

where $H_i^{(r,R)}$ satisfies (3.8a) *and* (4.2b) *taking into account remark* 3.10. *Probabilities $p_0, ..., p_N$ form the unique solution of the following system of linear equations:*

$$\sum_{i=0}^{N} p_i \frac{d^k}{dz^k} \{ K_i(z) H_i^{(r,R)}(z) - z^i \}\Big|_{z=z_s} = 0,\ k = 0,..., k_s - 1,\ s = 1,...,S, \tag{4.6b}$$

where z_s are R roots of $z^R - K(z)$ in the region $\bar{B}(0,1)\backslash\{1\}$ with their multiplicities k_s such that $\sum_{s=1}^{S-1} k_s = R - 1$, and $z_S = 0$ is of multiplicity $k_S = N - R$, and

$$\sum_{i=0}^{N} p_i \left(\rho_i - \rho + \bar{g}^{(i)} + \mathfrak{D}_y^{R-1} \left\{ \frac{\mathfrak{g}^{(i)}(y)}{(1-y)^2} \right\} \right) = R - \rho, \tag{4.6c}$$

where $\bar{g}^{(i)}$ is determined in (3.9b) *and* (3.3),

$$\rho_i = \lambda_i \alpha_i b_i. \tag{4.6d}$$

Proof. Formula (4.6a) follows from $P(z) = \sum_{i \in \Psi} p_i A_i(z)$ and (4.2-4.2a). It is easy to modify formula (4.6a) into

$$\sum_{i=N+1}^{\infty} p_i z^{i-N-1} = \frac{\sum_{i=0}^{N} p_i \{ K_i(z) H_i^{(r,R)}(z) - z^i \}}{z^{N+1} - z^{N+1-R} K(z)}, \tag{4.6e}$$

so that the function on the left-hand side of (4.6e) is analytic in $B(0,1)$ continuous on $\partial B(0,1)$.

According to theorem 4.4, for $\rho < R$, the function $z \mapsto z^R - K(z)$ has exactly R zeros in $\bar{B}(0,1)$ (counting with their multiplicities); all zeros located on the boundary $\partial B(0,1)$ (including 1), are simple. Therefore, the denominator in the right-hand side of (4.6e) must have exactly N roots in the region $\bar{B}(0,1)\backslash\{1\}$. This fact along with $(\boldsymbol{P},1) = 1$ (which yields (4.6c)) leads to equations (4.6b-4.6d).

The uniqueness of $\{p_0,...,p_N\}$ follows from the following considerations. Suppose that the system of equations (4.6b-4.6d) has another solution $p^* = \{p_i^*;\ i = 0,...,N\}$ which we substitute into (4.6a) to obtain the generating function $P^*(z)$. Then, $P^*(z)$ is analytic in $B(0,1)$ and continuous on $\partial B(0,1)$. Therefore, $\boldsymbol{P}^* = \{p_i^*; i \in \Psi\} \in (l^1, \|\cdot\|)$. Obviously, equations $P^*(z) = \sum_{i\in\Psi} p_i^* A_i(z)$ and $P^*(z) = \dfrac{\sum_{i=0}^{N} p_i^* \left\{ K_i(z) H_i^{(r,R)}(z) - z^i K(z) \right\}}{z^R - K(z)}$ are equivalent. The last equation is also equivalent to $\boldsymbol{P}^* = \boldsymbol{P}^* A$. Since p^* satisfies (4.6c) it follows that $(\boldsymbol{P}^*,1) = 1$. Thus, the system of equations $\boldsymbol{x} = \boldsymbol{x}A$, $(\boldsymbol{x},1) = 1$ has two different solutions in $(l^1, \|\cdot\|)$ which is impossible (*cf.* Gihman and Skorohod [22], theorem 15, p. 108). □

Below we give another version of theorem 4.6, where the corresponding formulas will be analytically less elegant but numerically are of greater advantage, especially for large N.

4.7 Theorem. *Given the condition $\rho < R$, $P(z)$ is determined by the following formula:*

$$P(z) = \frac{\sum_{i=0}^{R-1} p_i [z^R \{A_i(z) + H_i(z)\} - z^i K(z)]}{z^R - K(z)}, \tag{4.7a}$$

where $A_i(z)$ satisfies one of the expressions (4.2a-4.2c), $H_i(z) = \sum_{j=R}^{N} h_i^{(j)} z^{j-R} [K_i(z) - K(z)]$ *and $h_i^{(j)}$ are coefficients in the representations $p_j = \sum_{i=0}^{R-1} p_i h_i^{(j)}, j = R,...,N$. The unknown probabilities $p_0,...,p_{R-1}$ on the right-hand side of* (4.7a) *form a unique solution of the following system of R linear equations:*

$$\sum_{i=0}^{R-1} p_i \frac{d^k}{dz^k} \{A_i(z) + H_i(z) - z^i\}\Big|_{z=z_s} = 0,\ k = 0,...,k_s - 1,\ s = 1,...,S, \tag{4.7b}$$

$$\sum_{i=0}^{R-1} p_i \left(A_i(z) + H_i(z)]z^R - K(z)z^i \right)\Big|_{z=1} = R - \rho, \tag{4.7c}$$

where z_s are the roots of $z^R - K(z)$ in the region $\bar{B}(0,1)\backslash\{1\}$ with their multiplicities k_s such that $\sum_{s=1}^{S} k_s = R - 1$.

Proof. The statement of the theorem follows from theorem 4.6 if we take advantage of structural properties of the transition probability matrix A. It can be noticed that the matrix A a *positive essential N-homogeneous Δ_R-matrix* (in terms of [4]). Due to the specific features of this matrix we can uniquely express each of the probabilities $p_R,...,p_N$ on the right-hand side of

(4.8a) as a linear combination of the first probabilities $p_0,\ldots, p_{R-1}$:

$$p_j = \sum_{i=0}^{R-1} h_i^{(j)} p_i,\ j = R,\ldots,N. \tag{4.7d}$$

Taking into account these relations and proceeding as in the proof of formula (4.6d), we obtain (4.8a). The rest of the statement can be proved similar to theorem 4.6. □

Observe that, unlike formula (4.6b), that gives equations for finding N unknown probabilities, the right-hand side of (4.7a) contains only R unknown probabilities. This may be advantageous in computations, especially when $N \gg R$.

5. APPLICATIONS

5.1 Definitions.

(*i*) Let $\beta_j = E^j[T_1]$ and $\boldsymbol{\beta} = (\beta_j; j \in \Psi)^{\mathrm{T}}$. Then we will call the value $\boldsymbol{P\beta}$ the *mean service cycle of the system*, where $\boldsymbol{P}$ denotes the stationary probability distribution vector of the embedded queueing process Q_n.

(*ii*) Let $\boldsymbol{\alpha} = (\alpha_x; x \in \Psi)^{\mathrm{T}}$, $\boldsymbol{\lambda} = (\lambda_x; x \in \Psi)^{\mathrm{T}}$ and let $\boldsymbol{\rho} = \boldsymbol{\alpha}*\boldsymbol{\beta}*\boldsymbol{\lambda}$ be the Hadamard (entry-wise) product of vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. We call the scalar product $\boldsymbol{P\rho}$ the *intensity of the system.*

Observe that the notion of the "intensity of the system" (frequently called the *offered load* in queueing theory) goes back to the classical $M/G/1$ system, when $\boldsymbol{P\rho}$ reduces to $\rho = \lambda b$.

(*iii*) Define $l = \lim_{n\to\infty} E^x[inf\{Q_{\tau_{\nu_n}}, R\}]$ and call it the *mean (stationary) server load.*

5.2 Proposition. *Given the equilibrium condition $\rho < R$, the mean service cycle can be determined from the following expression:*

$$\boldsymbol{P\beta} = b + \sum_{j=0}^{N} p_j\left(b_j - b + \frac{1}{\lambda_j}\bar{\gamma}^{(j)}\right) \tag{5.2a}$$

Proof. Obviously, $\beta_j = b_j + \bar{\gamma}^{(j)}/\lambda_j$. The statement follows after elementary algebraic transformations. □

Using (5.2a) we similarly get

$$\boldsymbol{P\rho} = \rho + \sum_{i=0}^{N} p_i(\rho_i - \rho) + \sum_{i=0}^{r-1} p_i \bar{\jmath}^{(i)}. \tag{5.3}$$

5.4 Theorem. *Given the equilibrium condition $\rho < R$, the intensity of the system $\boldsymbol{P\rho}$ and the mean server load l coincide.*

Proof. Because of (2.2) we have

$$\begin{aligned} l = \sum_{i=0}^{\infty} p_i E^i[inf\{Q(\theta_0), R\}] &= \sum_{i=0}^{r-1} p_i E^i[S_{\nu_0} I_U \circ S_{\nu_0}] \\ &+ \sum_{i=0}^{r-1} p_i E^i[R\, I_{U^c} \circ S_{\nu_0}] + \sum_{i=r}^{R-1} i p_i + R\sum_{i=R}^{\infty} p_i. \end{aligned} \tag{5.4a}$$

Formula (4.6a) can be rewritten in the form

$$P(z) = \frac{\sum_{i=0}^{N} p_i\{z^R E^i[z^{Q_1}] - z^i E^{N+1}[z^{V_1}]\}}{z^R - E^{N+1}[z^{V_1}]}.$$

Then $(P, 1) = 1$ yields

$$\sum_{i=0}^{N} p_i\{E^i[Q_1] - \rho + R - i\} = R - \rho$$

or because of (2.3)

$$\sum_{i=0}^{r-1} p_i E^i[(S_{\nu_0} - R)I_{U^c} \circ S_{\nu_0}] + \sum_{i=R}^{N} p_i(i-R) + \sum_{i=0}^{N} p_i(\rho_i - \rho + R - i) = R - \rho.$$

The last expression allows modification of (5.3) into

$$P\rho = \sum_{i=0}^{r-1} p_i E^i[S_{\nu_0} I_U \circ S_{\nu_0}] + \sum_{i=r}^{R-1} i p_i + R\sum_{i=R}^{\infty} p_i + \sum_{i=0}^{r-1} p_i R P^i\{S_{\nu_0} \geq R\}$$

which, because of (5.4a), obviously equals l. □

Therefore, for the mean server load l we can use formula (5.3) which requires the knowledge of $p_0,\ldots,p_N$. The following formula for l is less friendly but it requires just $p_0,\ldots,p_{R-1}$.

5.5 Proposition. *The mean server load l can alternatively be obtained from the following formula:*

(5.5a)
$$l = R - \sum_{i=0}^{R-1} p_i \mathfrak{D}_y^{R-1}\left\{\frac{\mathfrak{G}^{(i)}(y)}{(1-y)^2}\right\}.$$

Proof. The statement directly follows from (5.3) and (4.6c). □

6. GENERAL QUEUEING PROCESS

In this section our main objective is the stationary distribution of the queueing process with continuous time parameter. Although this section is developed in a similar way as section 5 in Dshalalow [16] for the sake of consistency and better readability we include all necessary details. We need the following

6.1 Definitions.

(*i*) A stochastic process $\{\Omega, \mathcal{F}, (P^x)_{x\in\Psi}, Q(t);\ t \geq 0\} \to (\Psi, \mathfrak{B}(\Psi))$ with $\Psi \preceq \mathbf{N}$ is called *semi-regenerative* if

a) there is a point process $C = \sum_{n=0}^{\infty} \varepsilon_{T_n}$ on $\mathbf{R}_+$ such that $T_n \to \infty$ $(n\to\infty)$ and that each T_n is a stopping time relative to the canonic filtering $\sigma(Q_y; y \leq t)$,

b) the process Q has the locally strong Markov property at T_n, $n = 1,2,\ldots$, (see definition 4.2),

c) $\{Q(T_n + 0), T_n;\ n = 0,1,\ldots\}$ is a Markov renewal process.

(*ii*) Let (Q_n, T_n) be an irreducible aperiodic Markov renewal process with a discrete state space Ψ. Denote $\beta_x = E^x[T_1]$ as the mean sojourn time of the Markov renewal process in state

$\{x\}$ and let $\beta = (\beta_x; x \in \Psi)^T$. Suppose that the embedded Markov chain (Q_n) is ergodic and that P is its stationary distribution. We call $P\beta$ the *mean inter-renewal time.* Then we call the Markov renewal process *recurrent-positive* if its mean inter-renewal time is finite. An irreducible aperiodic and recurrent-positive Markov renewal process is called *ergodic.*

(*iii*) Let $\{Q_n, T_n\}$ be an ergodic Markov renewal process. For each x and j the following function $t \mapsto R^x(j,t) = E^x\left[\sum_{n=0}^{\infty} I_{\{j\}\times[0,t]} \circ (Q_n, T_n)\right]$, related to the Markov renewal process, is called a *Markov renewal function.* This function gives the expected number of entrances of the embedded Markov chain Q_n in state $\{j\}$ during time interval $[0,t]$, given that the process started from state $\{x\}$. Obviously, the process $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi}, \xi(t);\ t \geq 0\} \to (\mathbb{R}_+,\mathfrak{B}_+(\mathbb{R}_+))$ is the *minimal semi-Markov process associated with the Markov renewal process.* The process ξ is right continuous with almost every path as a simple function on any compact interval.

(*iv*) Let $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi}, Q(t);\ t \geq 0\} \to (\Psi, \mathfrak{B}(\Psi))$ be a semi-regenerative process relative to the sequence $\{T_n\}$ of stopping times. Introduce the probability

$$K_{jk}(t) = P^j\{Q(t) = k, T_1 > t\},\ j,k \in \Psi.$$

We will call the functional matrix $K(t) = (K_{jk}(t);\ j,k\in\Psi)$ the *semi-regenerative kernel.*

From the discussion in section 4 and from definition 6.1 (*i*), it follows that $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi}, Q(t);\ t \geq 0\} \to (\Psi, \mathfrak{B}(\Psi))$ is a semi-regenerative process with conditional regenerations at points T_n, $n = 0,1,\ldots$, $T_0 = 0$. By definition 6.1 (*ii*), $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi}, Q_n, T_n;\ n = 0,1,\ldots\}$ $\to$ $(\Psi\times\mathbb{R}_+,\ \mathfrak{B}(\Psi\times\mathbb{R}_+))$ is the associated Markov renewal process. Let $\mathcal{G}(t)$ denote the corresponding semi-Markov kernel. Under a very mild restriction to the probability distribution functions B_i, we can assume that the elements of $\mathcal{G}(t)$ are not step functions which would imply that $\{Q_n, T_n\}$ is aperiodic. By proposition 5.2, the mean service cycle $P\beta$, which is also the mean inter-renewal time of the Markov renewal process, is obviously finite. Therefore, following definition 6.1 (*iii*), the Markov renewal process is ergodic given the condition $\rho < R$.

It also follows that the jump process $\{\Omega,\mathfrak{F},(P^x)_{x\in\Psi}, \xi(t);\ t \geq 0\} \to \Psi$, defined in section 2, is the minimal semi-Markov process associated with Markov renewal process $\{Q_n, T_n\}$ and therefore, following definition 2.1, the input process Z^ξ is a compound Poisson process modulated by the semi-Markov process ξ.

6.2 Notation. Let

$$\delta_{xs}(t) = P^x\{Z^\xi([0,t]) = s \mid T_1 > t\}\ . \tag{6.2a}$$

Then, given that $\xi(0) = x$ and because Z^ξ is not modulated by a new value of ξ, the input process takes on value Z_x (introduced in definition 2.1 (*ii*)). Therefore, we have

$$\delta_{xs}(t) = P\{Z_x([0,t]) = s\}\ . \tag{6.2b}$$

□

Let $K(t) = (K_{jk}(t); j, k \in \Psi)$ be the semi-regenerative kernel (see definition 6.1 (*iv*)). The following statement holds true.

6.3 Lemma. *The semi-regenerative kernel satisfies the following equations:*

(6.3a)

$$K_{jk}(t) = \begin{cases} \delta_{j,k-j}(t), & 0 \le j \le k < r \\ \sum_{s=r-j}^{k-j} \int_0^t \varphi_j(s+j, t-u)\, \delta_{j,k-j-s}(u)(1 - B_j(u))\, du, & 0 \le j < r \le k \\ \delta_{j,k-j}(t)[1 - B_j(t)], & r \le j \le k \\ 0, & 0 \le k < j, \end{cases}$$

where δ_{jk} is as defined in (6.2a) *or* (6.2b) *and φ_j denotes the density of the joint probability distribution function of the random variable S_{ν_0} and the instant θ_0 of the first passage time of level r by the queueing process $\{Q(t)\}$ (defined by* (2.2)).

Proof. The above assertion follows from direct probability arguments. □

Now we are ready to apply the Main Convergence Theorem to the semi-regenerative kernel in the form of corollary 6.5.

6.4 Theorem (The Main Convergence Theorem, *cf.* Çinlar [9]). *Let $\{\Omega, \mathfrak{F}, (P^x)_{x \in \Psi}, Q(t); t \ge 0\} \to (\Psi, \mathfrak{B}(\Psi))$ be a semi-regenerative stochastic process relative to the sequence $\{t_n\}$ of stopping times and let $K(t)$ be the corresponding semi-regenerative kernel. Suppose that the associated Markov renewal process is ergodic and that the semi-regenerative kernel is Riemann integrable over $\mathbb{R}_+$. Then the stationary distribution $\boldsymbol{\pi} = (\pi_x; x \in \Psi)$ of the process $\{Q(t)\}$ exists and it is determined from the formula:*

(6.4a) $$\pi_k = \frac{1}{P\beta} \sum_{j \in \Psi} p_j \int_0^\infty K_{jk}(t)dt, \quad k \in \Psi.$$

6.5 Corollary. *Denote $H = (h_{jk}; j,k \in \Psi) = \int_0^\infty K(t)dt$ as the integrated semi-regenerative kernel, $h_j(z)$ the generating function of jth row of matrix H, $\boldsymbol{h}(z) = (h_j : j \in \Psi)^{\mathrm{T}}$ and $\pi(z)$ as the generating function of vector $\boldsymbol{\pi}$. Then the following formula holds true.*

(6.5a) $$\pi(z) = \frac{P\boldsymbol{h}(z)}{P\beta}.$$

Proof. From (6.4a) we get an equivalent formula in matrix form, $\boldsymbol{\pi} = \frac{PH}{P\beta}$. Finally, formula (6.5a) is the result of elementary algebraic transformations. □

6.6 Theorem. *Given the equilibrium condition $\rho < R$ for the embedded process $\{Q_n\}$, the stationary distribution $\boldsymbol{\pi} = (\pi_x; x \in \Psi)$ of the queueing process $\{Q(t)\}$ exists; it is independent of any initial distribution and is expressed in terms of the generating function $\pi(z)$ of $\boldsymbol{\pi}$ by the following formulas:*

(6.6a) $$P\beta\pi(z) = d(z)P(z) + \sum_{i=0}^{N} p_i\left[\frac{1}{\lambda_i}K_i(z)G_i(z) + z^i(d_j(z) - d(z))\right], \text{ with}$$

(6.6b) $$d_j(z) = \frac{1 - K_j(z)}{\lambda_j(1 - a_j(z))},$$

where $P(z)$ is the generating function of P, $P\beta$ is determined in proposition 5.2, *$G_i(z)$ is determined in* (3.5a) (*taking into account Remark 3.7*), *and $d(z)$ is defined as $d_j(z)$ with all subscripts dropped.*

Proof. Recall that the Markov renewal process $\{Q_n, T_n\}$ is ergodic if $\rho < R$. By corollary 6.5 the semi-regenerative process $\{Q(t)\}$ has a unique stationary distribution π provided that $\rho < R$. From (6.3a) we can see that the semi-regenerative kernel is Riemann integrable over $\mathbb{R}_+$. Thus, following corollary 6.5 we need to find the integrated semi-regenerative kernel H (which is done with routine calculus) and then generating functions $h_j(z)$ of all rows of H. First we find that

(6.6c) $$\sum_{p=i}^{\infty} z^p \int_0^{\infty} \delta_{i,p-i}(u)[1 - B_i(u)]\,du = z^i d_i(z)\ .$$

Then it follows that

(6.6d) $$h_i(z) = z^i \frac{1}{\lambda_i}\mathfrak{D}_x^{r-i-1}\left\{\frac{1}{(1-x)(1-a_i(xz))}\right\} + d_i(z)\mathfrak{G}_i^{(r)}(z),\ 0 \le i < r\ ,$$

where $\mathfrak{G}_i^{(r)}(z)$ denotes the tail of the generating function $\mathfrak{G}^{(i)}(z)$ summing its terms from r to ∞. However, it is easy to show that $\mathfrak{G}_i^{(r)}(z)$ and $\mathfrak{G}^{(i)}(z)$ coincide. Then it appears that

(6.6e) $$h_i(z) = z^i d_i(z)\,,\ i \ge r,$$

where the index i can be dropped for all i exceeding N, in accordance with assumption (AS) made in section 4. Formula (6.6a) now follows from corollary 6.5, equations (6.6c-6.6e), (3.5a), (3.6a) and some algebraic transformations. □

7. EXAMPLES AND SPECIAL CASES

In one of the first examples we drop the modulation of the input preserving all other special features of the system.

7.1 Proposition. *In the bulk queueing system with no modulation of the input the generating*

function $\pi(z)$ of π can be derived from the following formula:

(7.1a) $$\frac{l[1 - a(z)]}{\alpha}\pi(z) = P(z)(1 - z^R) + \sum_{i=0}^{R-1} p_i K_i(z)\mathfrak{D}_y^{R-1}\left\{\frac{z^R\mathfrak{G}^{(i)}(y) - \mathfrak{G}^{(i)}(yz)}{1 - y}\right\}.$$

Proof. Formula (7.1a) follows from (6.6a) and (4.6a) after noticing that

(7.1b) $$\mathfrak{D}_y^{R-1}(y^i f(y)) = 0 \text{ for all } i \ge R,$$

where f is any function analytic at the origin. □

Formula (7.1a) may look unfriendly but in the way it is presented it yields a number of elegant special cases.

7.2 Examples.

(*i*) Let $r = R$. Then the sum in the right-hand side of (7.1a) vanishes by the reason observed in (7.1b). Obviously, in this case l reduces to r finally yielding an elegant relation between $\pi(z)$ and $P(z)$:

$$\pi(z) = \frac{\alpha(1-z^r)}{r[1-a(z)]} P(z)\,. \tag{7.2a}$$

(*ii*) Suppose that arriving groups are distributed geometrically. In other words, let $a_i(z) = a(z) = pz(1-qz)^{-1}$. Then from (3.2a) we have

$$\mathfrak{g}^{(i)}(z) = z^r p(1-qz)^{-1} \text{ for } i < r \tag{7.2b}$$

$$\mathfrak{g}^{(i)}(z) = z^i \text{ for } i \geq r \tag{7.2c}$$

and thus $E^i[z^{S_{\nu_1} - r}] = p(1-qz)^{-1}$, yielding that the random variable $S_{\nu_n} - r$ is "memoryless" in this special case.

Now substitute (7.2b) and (7.2c) into (7.1a) and get

$$\pi(z)\frac{1-z}{1-qz} = \frac{1}{pl} P(z)(1-z^R) + \frac{1}{l} \sum_{i=0}^{R-1} p_i K_i(z) W_i(z), \tag{7.2d}$$

where

$$W_i(z) = \begin{cases} \dfrac{z^R(1-q^{R-r-1})}{1-q} - \dfrac{z^r[1-(qz)^{R-r-1}]}{1-qz}, & 0 \leq i < r \\ \dfrac{1}{p}(z^R - z^i), & r \leq i < R\,. \end{cases} \tag{7.2e}$$

The mean server load l can be evaluated from formula (5.5a) that leads to

$$l = R - \sum_{i=r}^{R-1} p_i(R-i) - \frac{(R+1)p + q^{R+1} - q}{p^2} \sum_{i=0}^{r-1} p_i\,. \tag{7.2f}$$

(*iii*) In the condition of (*ii*) assume additionally that $r = R-1$. Then the upper expression in (7.2e) vanishes reducing (7.2d) to

$$\pi(z)\frac{pl}{1-qz} = P(z)\frac{1-z^{r+1}}{1-z} - p_r K_r(z) z^r. \qquad \square$$

In the next situation we suppress bulk arrivals however preserving the modulation.

7.3 Theorem. *In the multilevel controlled queueing system with (r, R)-queue length dependent service delay discipline with an orderly modulated Poisson stream the generating function $\pi(z)$ satisfies the following relation:*

(7.3a) $$\lambda P\beta(1-z)\pi(z) = [1-K(z)]P(z) + \sum_{i=0}^{N} p_i[z^i K(z) - z^{r+(i-r)^+} K_i(z)], \quad \text{where}$$

(7.3b) $$P(z) = \frac{\sum_{i=0}^{N} p_i\{z^{R+(i-R)^+} K_i(z) - z^i K(z)\}}{z^R - K(z)}.$$

Probabilities $p_0, \ldots, p_N$ *form the unique solution of the system* (4.6b) *and* (4.6c), *where* $H_i^{(r,R)}(z) = z^{(i-R)^+}$. *Equation* (4.6d) *is reduced to*

(7.3c) $$\sum_{i=0}^{N} p_i[\rho_i - \rho + (R-i)^+] = R - \rho.$$

Proof. Formula (7.3a) follows from (6.6a) and (6.6b) after some algebraic transformations.

Formula (7.3b) follows from (3.8b), (4.2b), (4.2d) and (4.6a) and because $\mathfrak{g}^{(i)}(z)$ reduces to

(7.3d) $$\mathfrak{g}^{(i)}(z) = \begin{cases} z^r, & i < r \\ z^i, & i \geq r. \end{cases}$$

Equation (7.3c) is due to $\bar{\mathfrak{g}}^{(i)} = (r-i)^+$ valid for this special case. □

7.4 Examples.

(*i*) In the condition of theorem 7.3 the mean server load is defined by

$$l = \sum_{i\in\Psi} p_i \, min\{R, max\{r, i\}\} = R\sum_{i=R}^{\infty} p_i + r\sum_{i=0}^{r-1} p_i + \sum_{i=r}^{R-1} ip_i$$

that also agrees with formula (5.5a) to which this special case is applied.

(*ii*) In the condition of theorem 7.3 by dropping the modulation of the input we have from (7.1a) and (7.3d) that

$$l(1-z)\pi(z) = P(z)(1-z^R) + \sum_{i=0}^{R-1} p_i K_i(z)(z^R - z^{r_i}), \quad \text{where}$$

$$r_i = r, \; i < r \text{ and } r_i = i, \; i \geq r.$$

(*iii*) If the input is a stationary compound Poisson process (i.e. nonmodulated) then its intensity is $\alpha\lambda$, which is also the mean number of arriving units per unit time. In the case of a modulated input process its intensity is no longer a trivial fact. We define the intensity of any random measure Z by the formula $\kappa = lim_{t\to\infty} \frac{1}{t} E^x[Z([0,t])]$. We will apply the formula from theorem 7.5 (Dshalalow [9]) for more general Poisson process modulated by a semi-Markov process:

$$\kappa = \frac{P\rho}{P\beta},$$

where by theorem 5.4 $P\rho = l$ and $P\beta$ satisfies (5.2a). Thus we have that:

(7.4a) $$\kappa = \frac{l}{P\beta}.$$

(*iv*) By virtue of obvious probability arguments we can derive the probability density function of an idle period in the steady state:

$$u \mapsto \frac{\sum_{i=0}^{r-1} p_i \gamma^{(i)}(\theta, 1)}{\sum_{i=0}^{r-1} p_i}.$$

The mean value of the idle period $\mathfrak{I}$ in the steady state is then

$$\mathfrak{I} = \frac{\sum_{i=0}^{r-1} p_i \bar{\gamma}^{(i)} \frac{1}{\lambda_i}}{\sum_{i=0}^{r-1} p_i} . \tag{7.4b}$$

(v) Formula (7.4b) and theorem 6.6 allow to derive the mean busy period $\mathfrak{B}$ in the equilibrium. Clearly $\sum_{i=0}^{r-1} \pi_i$ is the probability that the server idles. On the other hand, it also equals $\frac{\mathfrak{I}}{\mathfrak{I}+\mathfrak{B}}$. Thus we have

$$\mathfrak{B} = \frac{\mathfrak{I} \sum_{n=r}^{\infty} \pi_n}{\sum_{i=0}^{r-1} \pi_i} .$$

7.5 Theorem (Dshalalow [13]). *The stationary rate of the input flow averaged over the infinite horizon is determined from the following formula:*

$$\kappa = \lim_{t\to\infty} \frac{E^x[Z^{\xi}([0,t])]}{t} = \frac{P\rho}{P\beta} . \tag{7.5a}$$

Ergodic Theorems for Some Functionals of Input and Output Processes.

One of the goals of this section is to find $\lim_{t\to\infty} \frac{E^x[Q(t)]}{t}$. By a direct computation, it can be shown that for $\rho < R$ the value of $\lim_{t\to\infty} E^x[Q(t)]$ is a function of the second moment of a service time that need not be finite. In the latter case, it is not obvious with what speed $\lim_{t\to\infty} E^x[Q(t)]$ gets to infinity. We will show that, even if it diverges, it gets slower to infinity than with the speed.

Let B be a Borel set on $\mathbb{R}_+$. Denote $S(B)$ the total number of customers completely processed on the time-set B.

7.6 Theorem. *For $\rho < R$ the output rate defined $\mathcal{O} = \lim_{t\to\infty} \frac{E^x[S([0,t])]}{t}$ equals the ratio $\frac{P\rho}{P\beta}$.*

Proof. Obviously

$$\begin{aligned} S([0,t]) = {} & \sum_{j\in\Psi} \sum_{n=0}^{\infty} inf\{Q(\theta_n \circ Q_n), R\} I_{\{j\}\times[0,t]} \circ \{Q_n, T_n\} \\ & - \sum_{j\in\Psi} inf\{Q(\theta_{C([0,t])} \circ \xi(t)), R\} I_{\{j\}} \circ \xi(t) . \end{aligned} \tag{7.6a}$$

The sum in the second line of equation (7.6a) gives the total number of units being in service but not completely processed by time t. Clearly, this sum is majorated by R for every $\omega \in \Omega$. To find the output rate first observe that (7.6a) is reduced to

$$\begin{aligned} S([0,t]) = {} & \sum_{j\in\Psi} \sum_{n=0}^{\infty} inf\{Q(\theta_0(j)), R\} I_{\{j\}\times[0,t]} \circ \{Q_n, T_n\} \\ & - \sum_{j\in\Psi} inf\{Q(\theta_{C([0,t])} \circ \xi(t)), R\} I_{\{j\}} \circ \xi(t) , \end{aligned} \tag{7.6b}$$

since obviously the random variables $\theta_n \circ Q_n$, $n = 0,1,...,$ are independent and identically

distributed if given on the trace σ-algebra $\mathfrak{F}_t \cap \left[\bigcup_{n=0}^{\infty} Q_n^{-1}(\{j\})\right]$ (where $\mathfrak{F}_t$ is the canonic filtering induced by the process $\{Q(t)\}$). The latter enables one to evaluate the functional $E^x[S([0,t])]$ by using the independence of $Q(\theta_0(j))$ and $I_{\{j\}\times[0,t]} \circ \{Q_n, T_n\}$. Applying the monotone convergence theorem and in the light of definition 6.1 (*iii*) we have

$E^x[S([0,t])] = \sum_{j\in\Psi} E^j[inf\{Q \circ \theta_0\}]R^x(j,t) - E^x[\sum_{j\in\Psi} inf\{Q(\theta_{C([0,t])} \circ \xi(t)), R\} I_{\{j\}} \circ \xi(t)]$.

Since $E^x[\sum_{j\in\Psi} inf\{Q(\theta_{C([0,t])} \circ \xi(t)), R\} I_{\{j\}} \circ \xi(t)] \leq R$ for all $t \geq 0$ which simplifies the output rate to

$$\mathcal{O} = \lim_{t\to\infty} \sum_{j\in\Psi} E^j[inf\{Q \circ \theta_0\}] \frac{R^x(j,t)}{t}$$

finally yielding from theorem A.1 (*ii*) (see Appendix) that

$$\mathcal{O} = \frac{l}{P\beta}. \tag{7.6c}$$

Finally, by theorem 5.4 the mean server load and the intensity of the system coincide and this proves the theorem. □

Since Z^ξ is the input process modulated by the semi-Markov process ξ we can use formula (7.5a) in theorem 7.5 which gives the mean input rate κ of the modulated semi-Markov process Z^ξ. From (7.6a) and (7.6c) it therefore follows that $\kappa = \mathcal{O}$. This is to be expected in most of the systems thereby proving valid one of the conservation laws: "*In an ergodic stochastic system the input and output rates are equal*".

7.7 Corollary. *For $\rho < R$ the expected number of units in the system in equilibrium is either finite or diverges slower than with the unit speed.*

Proof. Since the number of units in the system at time t is $Q(t) = Q(0) + Z^\xi([0,t]) - S([0,t])$ the statement follows by theorems 7.5 and 7.6. □

7.8 Example. As an application of the ergodic theorems 7.6 and A.1 (see appendix), we consider the following optimization problem. Let c_1, c_2, c_3, c_4, w be real-valued Borel-measurable functions that represent the following cost rates and functionals:

$c_1(k)$ denotes the total expenses due to the presence of k customers in the system per unit time. Then $F_1[c_1,Q](x,t) = E^x[\int_0^t c_1(Q(u))\,du]$ gives the expected expenses due to the presence of all customers in the system in time interval $[0,t]$ given that initially x units were present. By Fubini's theorem and theorem A.1 (*ii*) we have $\lim_{t\to\infty} \frac{1}{t} F_1[c_1, Q](x,t) = \sum_{j\geq 0} c_1(j)\pi_j = \boldsymbol{\pi}\mathbf{c_1}$ as the expected cost rate due to the presence of all units in the system, where $\mathbf{c_1} = (c_1(0), c_1(1),...)^T$.

$c_2(j)$ denotes the expenses for the service act of type j per unit time [observe that the decision to "apply a certain distribution function B_j" when the system accumulated j units, will be affected by the cost function c_2 that is usually inverse proportionally to the service rates].

Thus, $F_2[c_2, \xi](x,t) = E^x[\int_0^t c_2(\xi(u))du]$ is the expected cost of all service acts in time interval $[0,t]$ given that initially x units were present in the system. By Fubini's Theorem and theorem A.1 (*iii*) we have $\lim_{t\to\infty} \frac{1}{t} F_2[c_2, \xi](x,t) = \frac{P\beta * c_2}{P\beta}$ (where $\beta * c_2$ denotes the Hadamard product of β and $c_2 = (c_2(0), c_2(1),...)^T$) as the expected cost rate for all service acts over infinite horizon.

c_3 is a real-valued scalar denoting the penalty for each interruption of continuously operating service per unit time [by an interruption of continuously operating server we understand each "entrance" of the server in idle period when it has to wait for the queue level to reach or exceed r]. Then, $c_3 \sum_{i=0}^{r-1} R^x(i,t)$ gives the expected expenses for $\sum_{i=0}^{r-1} R^x(i,t)$ number of idlenesses of the server in time interval $[0,t]$. By theorem A.1 (*i*) the penalty rate for each entrance in an idle period equals $\lim_{t\to\infty} \sum_{i=0}^{r-1} c_3 \frac{1}{t} R^x(i,t) = \sum_{i=0}^{r-1} \frac{c_3 p_i}{P\beta}$. Observe that a necessity to penalize the system for service interruptions has a good reason to reduce warm-up expenses. Since our service time distribution functions are arbitrary and all of them may be different it includes, as an option, a warm-up time prior to the service, so that B_i may be given in the form of a convolution of two probability distribution functions.

c_4 denotes the penalty for a unit time to spend idle by the server. Since the expected time the server idles "on" a set $B \in \mathfrak{B}_+$ is

$$E^x[\sum_{i=0}^{r-1} \int_B I_{\{i\}} \circ Q(u)\, du] = \sum_{i=0}^{r-1} \int_B P^x\{Q(u) = i\}\, du,$$

we have by theorem A.1 (*ii*) that

$$\lim_{t\to\infty} \frac{1}{t} E^x[\sum_{i=0}^{r-1} \int_0^t I_{\{i\}} \circ Q(u)\, du] = \sum_{i=0}^{r-1} \pi_i$$

giving $c_4 \sum_{i=0}^{r-1} \pi_i$ as the penalty rate for the server to idle per unit time averaged over infinite horizon.

w denotes the reward for each completely processed unit per unit time. By theorem 5.1, the expected gain of the system per unit time is $\lim_{t\to\infty} w \frac{1}{t} E^x[S([0,t])] = w \frac{P\rho}{P\beta} = \frac{wl}{P\beta}$.

Finally, the objective function Φ is then

$$\Phi = \frac{1}{P\beta}\left[wl - P\beta * c_2 - c_3 \sum_{i=0}^{r-1} p_i\right] - c_4 \sum_{i=0}^{r-1} \pi_i - \pi c_1.$$

APPENDIX

A.1 Theorem. *The below formulas hold true for the functionals of the stochastic processes defined in* 5.1(*i*) *and* 6.1(*iii*):

(*i*) $\lim_{t\to\infty} \frac{1}{t} R^x(j,t) = \frac{p_j}{P\beta}$

(*ii*) $\lim_{t\to\infty} \frac{1}{t} \int_0^t P^x\{Q(u) = j\}\, du = \pi_j$

(*iii*) $\lim_{t\to\infty} \frac{1}{t} \int_0^t P^x\{\xi(u) = j\}\, du = \frac{p_j b_j}{P\beta}$.

Proof.

(*i*) Let $\delta_0 = inf\{n > 0\colon Q_n = j,\ Q_0 = x\}$, $\delta_k = inf\{n > \delta_{k-1}\colon Q_n = j,\ Q_0 = x\}$. Then $\{T_{\delta_k};\ k \geq 0\}$ is a delayed renewal process (embedded in the point process $\{T_n\}$). Clearly, $\{T_{\delta_k};\ k \geq 0\}$ is recurrent if and only if $\rho < R$. Then it follows from Çinlar [9] and due to $\lim_{t\to\infty} P\{\delta_0 \leq t\} = 1$ for $\rho < R$ that

$$\lim_{t\to\infty} \frac{R^x(j, t)}{t} = \mu_j ,$$

where μ_j is the reciprocal of the mean time between two subsequent returns of Q_n to state $\{j\}$. On the other hand, from Markov renewal theory it is known that

$$\mu_j = \frac{p_j}{P\beta}$$

and the statement (*i*) then follows.

(*ii*) By Çinlar [9] we have

$$P^x\{Q(u) = k\} = \sum_{j\geq 0} \int_0^u R^x(j, ds)\, K_{jk}(t)$$

yielding that

$$\int_0^t P^x\{Q(u) = k\}\, du = \sum_{j\geq 0} R^x(j, \cdot) * g(t) ,$$

where $g(t) = \int_0^t K_{jk}(v)\, dv$ is a non-decreasing continuous function and symbol "$*$" denotes the convolution operator. Then, it follows that

$$\lim_{t\to\infty} \frac{R^x(j, \cdot) * g(t)}{R^x(k, t)} = \frac{p_j}{p_k} \int_0^\infty K_{jk}(u)\, du .$$

Now applying (*i*) and formula (6.4a) we finally obtain

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t P^x\{Q(u) = k\}\, du = \lim_{t\to\infty} \frac{\sum_{j\geq 0} R^x(j, \cdot) * g(t)}{R^x(k, t)} \frac{R^x(k, t)}{t} = \pi_k .$$

(*iii*) The statement follows directly from Çinlar [9]. □

Acknowledgement. The authors are very grateful to Professor Lajos Takács for his kind advise and helpful comments.

REFERENCES

[1]. Abolnikov, L., A single-level control in Moran's problem with feedback, *Oper. Res. Letters*, **2**, No.1, 16-19, 1983.

[2]. Abolnikov, L. and Dshalalow J., Feedback queueing systems; duality principle and optimization, *Autom. and Remote Control* (Izv. Akad. Nauk SSSR.), **39**, No.1, 11-20, 1978.

[3]. Abolnikov, L. and Dshalalow, J., A first passage problem and its applications to the analysis of a class of stochastic models, *Journ. Appl. Math. Stoch. Anal.*, **5**, No. 1, 83-98, 1992.

[4]. Abolnikov, L. and Dukhovny A., Markov chains with transition delta-matrix: ergodicity conditions, invariant probability measures and applications, *Journ. Appl. Math. Stoch. Anal.*, **4**, No. 4, 335-355, 1991.

[5]. Arnold, B. and Groenevald, R., On excess life in certain renewal processes, *Journ. Appl. Probab.*, **18**, 379-389, 1981.

[6]. Bachary, E., Kolesar, P., Multilevel bulk service queue, *Op. Res.*, **20**, No.2, 1972.

[7]. Chaudhry, M. and Templeton, J., *A First Course in Bulk Queues*, John Wiley and Sons, New York, 1983.

[8]. Chaudhry, M., Madill, B. and Briére, G., Computational Analysis of Steady-State Probabilities of $M/G^{a,b}/1$ and related nonbulk queues, *Queueing Systems*, **2**, 93-114, 1987.

[9]. Çinlar, E., *Introduction to Stochastic Processes*, Prentice Hall, Englewood Cliffs, N.J., 1975.

[10]. Cohen, J., On up- and downcrossings, *Journ. Appl. Probab.*, **14**, 405-410, 1977.

[11]. Deb, R., Serfoso, R., Optimal control of batch service queues, *Advances in Appl. Probab.*, **5**, 340-361, 1973.

[12]. Delbrouck, L., A feedback queueing system with batch arrivals, bulk service and queue-dependent service time, *Journ. Assoc. Comp. Mach.*, **17**, No.2, 314-323, 1970.

[13]. Dshalalow, J., On modulated random measures. *Journ. Appl. Math. Stoch. Anal.*, **4**, No.4, 305-312, 1991.

[14]. Dshalalow, J., A single-server queue with random accumulation level, *Journ. Appl. Math. Stoch. Analysis*, **4**, No. 3, 203-210, 1991.

[15]. Dshalalow, J., Single-server queues with controlled bulk service, random accumulation level, and modulated input, to appear in *Stoch. Analysis and Applications.*

[16]. Dshalalow, J., On a first passage problem in general queueing system with multiple vacations, *Journ. Appl. Math. Stoch. Anal.*, **5**, No. 2, 177-192, 1992.

[17]. Dshalalow, J. and Tadj, L., A queueing system with a fixed accumulation level, random server capacity, and capacity dependent service time, *Intern. Journ. Math. and Math. Sciences*, **15**, No. 1, 189-194, 1992.

[18]. Dshalalow, J. and Tadj, L., On a multiple control queue with a bilevel service delay policy and state dependent random server capacity, submitted to *Probab. in the Engin. Inform. Sciences.*

[19]. Dshalalow, J. and Tadj, L., On applications of first excess level random processes to queueing systems with random server capacity and capacity dependent service time, to appear in *Stoch. and Stoch. Reports.*

[20]. Dynkin, E., Some limit theorems for sums of independent random variables with infinite mathematical expectations, *Izv. Akad. Nauk SSSR*, Ser. Math. **19**, 247-266, 1955 [or in *Selected Translations in Mathematical Statistics and Probability*, IMS and AMS, **1**, 171-189, 1961].

[21]. Federgruen, A., Tijms, H.C., Computation of the stationary distribution of the queue size in an $M \mid G \mid 1$ queueing system with variable service rate, *Journ. Appl. Probab.*, **17**, 515-522, 1980.

[22]. Gihman, I. and Skorohod, A., *The theory of stochastic processes*, Vol. I, Springer-Verlag, 1974.

[23]. Neuts, M.F., A general class of bulk queues with Poisson input, *Ann. Math. Stat.*, 759-770, 1967.

[24]. Shanthikumar, J.G., Level crossing analysis of priority queues and a conservation identity fro vacation models, *Nav. Res. Loqist.*, **36**, 797-806, 1989.

[25]. Takács, L., On fluctuations of sums of random variables, *Advances in Mathematics*, **2**, 45-93, 1978.